Retail-Giant Sales and Quantity Forecasting Case Study

PRESENTED BY RAMA MISHRA

Data Understanding

Given is the data for 48 months consisting of 6 attributes.

From the business understanding, following attributes were of interest.

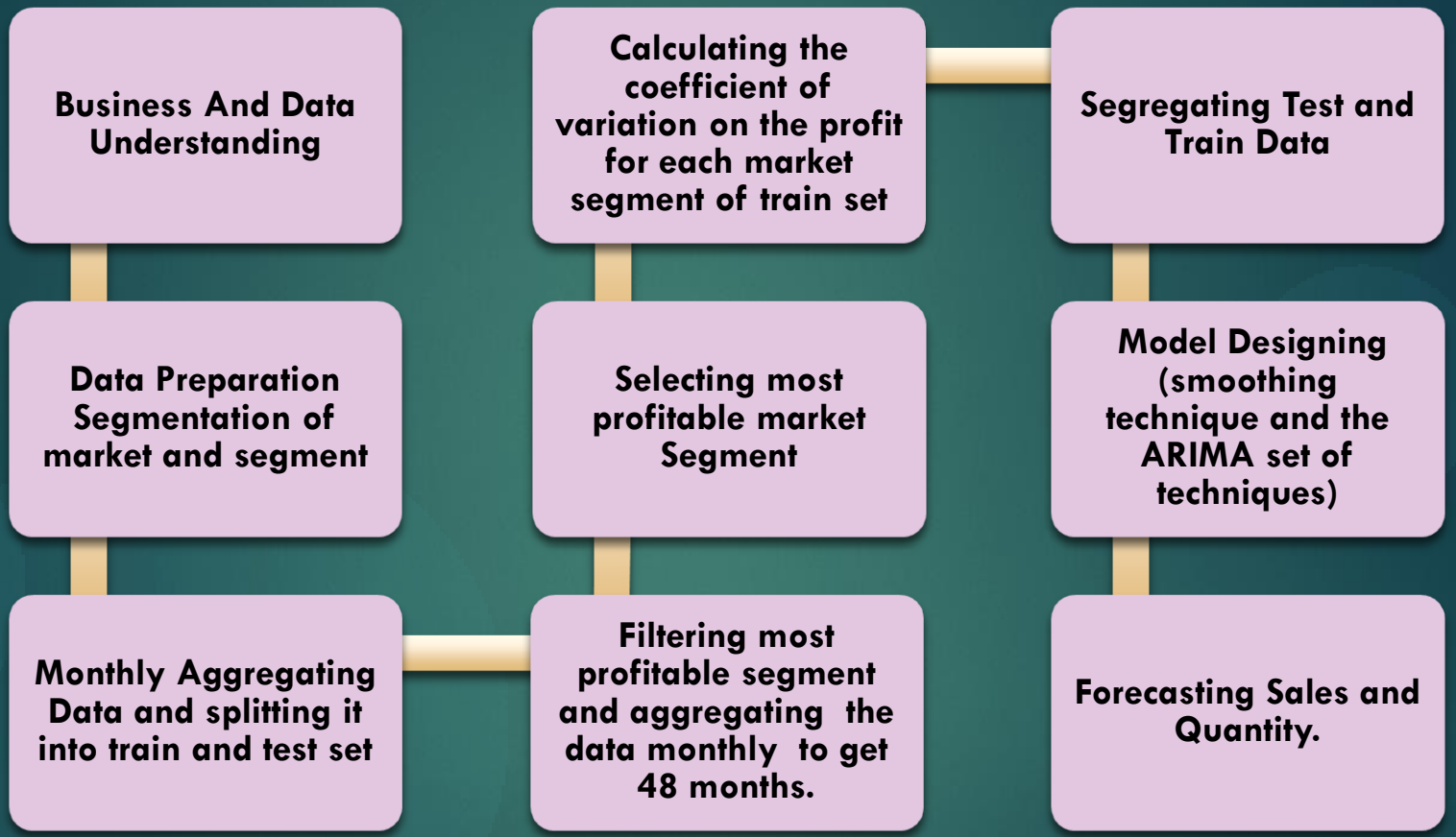
The store caters to 7 different market segments and in 3 major categories. Need to subset data into 21 (7*3) buckets for analysis.

- Market
- Customer Segment
- Sales
- Quantity
- Profit
- Order Date

Problem Statement & Objective

- ✓ Global Mart is having worldwide operations.
- ✓ Takes orders and delivers across the globe and deals with all the major product categories - consumer, corporate & home office.
- ✓ Wants to forecast the sales and the demand for the next 6 months to help manage the revenue and inventory accordingly.
- ✓ The store caters to 7 different market segments and in 3 major categories.
- ✓ Find out the most profitable and consistent segments and forecast the sales and demand for these segments.

Problem Solving



Data Preparation

Checking for NA values in desired attributes

Creating 21 buckets for each of the seven markets and their corresponding three segments.

Aggregation of Data in a monthly form

Coefficient of Variation Calculation for each bucket.

Problem solving methodology

Data Preparation

- Created 21 Market-Segment buckets.
- Aggregated buckets by Sales, Profit, Quantity.
- Splitted the data into train(42) and test(6).
- Calculated Coefficient of Variation(CoV) for train data(42).
- Selected the most profitable market_segment (APAC Consumer) based on least CoV value.
- Filter the data with most profitable market_segments and created the monthly aggregated data for 'Order Date' and 'Sales'.

Modeling using smoothing techniques

- From the monthly aggregated data for 'Order Date' and 'Sales' of APAC Consumer. Splitted the data into train & test sets of size 42 & 6 months.
- Builded the model by using smoothing techniques:-
- 1) Simple exponential smoothing.
- 2) Holt's exponential smoothing.
- 3) Holt-Winters' exponential smoothing (both additive and multiplicative techniques)

Modeling using ARIMA set of techniques

- Converted data into stationary by using box-cox transformation and differencing method.
- Builded the model by using ARIMA set of techniques:-
  1) AR model
  2) MA model
  3) ARMA model
  4) ARIMA model
  5) SARIMA model
- ARIMAX and SARIMAX model (can not be applied, as there is no exogenous variable in the data).

Forecasting

- Holt-Winters' exponential smoothing and SARIMA model were able to predict the forecast future 6 months sales & quantity from January 2011 to December 2014 closer to the actual values.

Data collection (Time series data)

- Import required packages and Import and read the data
- Convert the "Order Date" column into date- time format

```
# Import the NumPy and Pandas packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
# Supress unnecessary warnings
import warnings
warnings.filterwarnings("ignore")
```

Import time series data: Global2

```
# Read the dataset
Retail = pd.read_csv('Global2.csv')
# Changing the 'Order Date' column to datetime format
Retail['Order Date']= pd.to_datetime(Retail['Order Date'], format='%d-%m-%Y')
Retail.head()
```

	Order Date	Segment	Market	Sales	Quantity	Profit
0	2012-07-31	Consumer	US	2309.650	7	762.1845
1	2013-02-05	Corporate	APAC	3709.395	9	-288.7650
2	2013-10-17	Consumer	APAC	5175.171	9	919.9710
3	2013-01-28	Home Office	EU	2892.510	5	-96.5400
4	2013-11-05	Consumer	Africa	2832.960	8	311.5200

Data Preparation

- Created 21 data subset buckets based on Market & Segment they belong.

- Creating a Pivot table for getting the month wise total sum of profit by each market segment.

	Order Date	Segment	Market	Sales	Quantity	Profit	Market_Segment
0	2012-07-31	Consumer	US	2309.650	7	762.1845	US_Consumer
1	2013-02-05	Corporate	APAC	3709.395	9	-288.7650	APAC_Corporate
2	2013-10-17	Consumer	APAC	5175.171	9	919.9710	APAC_Consumer
3	2013-01-28	Home Office	EU	2892.510	5	-96.5400	EU_Home Office
4	2013-11-05	Consumer	Africa	2832.960	8	311.5200	Africa_Consumer
...	...	...	...	...	...	...	...
51285	2014-06-19	Corporate	APAC	65.100	5	4.5000	APAC_Corporate
51286	2014-06-20	Consumer	US	0.444	1	-1.1100	US_Consumer
51287	2013-12-02	Home Office	US	22.920	3	11.2308	US_Home Office
51288	2012-02-18	Home Office	LATAM	13.440	2	2.4000	LATAM_Home Office
51289	2012-05-22	Consumer	LATAM	61.380	3	1.8000	LATAM_Consumer

51290 rows × 7 columns

```
grouping the data using pivot table command such that the index is Order_Date, Columns are "Market-Segment" and values is Pro
df= pd.pivot_table(data=Retail, values='Profit', index=pd.Grouper(key='Order Date', freq='1M'), columns='Market_Segment', aggfu
df
```

Market_Segment	APAC_Consumer	APAC_Corporate	APAC_Home Office	Africa_Consumer	Africa_Corporate	Africa_Home Office	Canada_Consumer	Canada_Corporate	C
Order Date									
2011-01-31	991.2825	11.5998	86.4423	475.683	219.096	856.710	3.12	5.70	
2011-02-28	1338.8688	4358.8254	-417.4128	1441.926	-490.551	820.302	23.31	NaN	
2011-03-31	3747.1632	1213.3386	923.7492	322.140	-586.716	67.320	335.55	NaN	
2011-04-30	3846.4746	71.0265	657.1080	292.122	776.691	500.136	55.08	NaN	
2011-05-31	3639.9423	2534.1672	-272.1717	110.004	241.338	34.926	77.97	NaN	
2011-06-30	4328.2596	1435.8294	3452.1018	-1290.639	-259.218	-774.801	7.50	40.08	
2011-07-31	1258.9617	525.4647	-42.0498	621.168	134.847	-506.562	47.28	528.33	
2011-08-31	775.8066	4070.5581	775.7616	232.917	915.885	1090.200	43.68	29.70	
2011-09-30	5181.4449	1952.4675	623.3595	-86.163	950.766	1345.740	134.01	10.17	
2011-10-31	6911.9970	5452.2429	1131.9597	612.942	-97.821	387.312	35.43	NaN	
2011-11-30	221.5539	2154.5313	6574.6911	-221.124	608.700	639.633	135.66	33.60	
2011-12-31	4004.3211	8126.9478	1384.9611	642.921	875.220	41.064	NaN	39.03	
2012-01-31	4810.1535	2224.6740	709.7085	762.894	-192.171	35.760	NaN	NaN	
2012-02-29	2967.1281	659.0961	335.0127	-352.278	131.235	44.823	NaN	122.46	
2012-03-31	146.4261	1108.3668	1381.6839	774.156	498.090	-99.330	6.03	NaN	
2012-04-30	2368.1721	1873.7997	-97.5930	347.982	306.720	467.070	27.00	NaN	
2012-05-31	6114.3312	4751.5857	3500.2146	-342.684	-57.648	224.304	52.65	322.32	
2012-06-30	4790.0052	4853.2323	124.5438	2357.850	-421.314	-144.864	1785.84	426.54	

Data Preparation

- Splitted the data into train(42) and test(6) sets.

```
# splitting the data into train and test
train_len= 42
train= df[0:train_len]
test= df[train_len:]
```

- Calculated Coefficient of Variation(CoV) using aggregated Profit for each Market-Segment for train sets using below formula:

$$CoV = std(Profit)/mean(Profit)$$

- Using CoV, Selected most profitable Market-Segments as APAC_Consumer
- Filtering the data so that only the most profitable market segment (APAC Consumer) is present for further prediction.

```
# finding the "Coefficient of Variation (CoV)" of 'Profit' for train data
cov=std/mean
cov
```

Market_Segment	
APAC_Consumer	0.603633
EU_Consumer	0.655334
LATAM_Consumer	0.688935
EU_Corporate	0.697702
APAC_Corporate	0.740799
LATAM_Corporate	0.890930
US_Corporate	1.039660
APAC_Home Office	1.061530
US_Consumer	1.108571
EU_Home Office	1.128192
Canada_Corporate	1.219189
US_Home Office	1.231887
LATAM_Home Office	1.359984
Africa_Consumer	1.446661
Canada_Consumer	1.497032
Africa_Corporate	1.685008
Africa_Home Office	2.013987
Canada_Home Office	2.245148
EMEA_Consumer	2.749927
EMEA_Home Office	6.140222
EMEA_Corporate	6.861820

```
APAC_Consumer= Retail[(Retail.Market_Segment== 'APAC_Consumer')]
APAC_Consumer= APAC_Consumer.sort_values(by='Order Date')
APAC_Consumer
```

	Order Date	Segment	Market	Sales	Quantity	Profit	Market_Segment
42055	2011-01-01	Consumer	APAC	55.2420	2	15.3420	APAC_Consumer
22951	2011-01-01	Consumer	APAC	120.3660	3	36.0360	APAC_Consumer
31869	2011-01-01	Consumer	APAC	113.6700	5	37.7700	APAC_Consumer
2709	2011-01-03	Consumer	APAC	912.4560	4	-319.4640	APAC_Consumer
47553	2011-01-03	Consumer	APAC	6.0060	1	0.5460	APAC_Consumer
...	...	...	...	...	...	...	...
39629	2014-12-30	Consumer	APAC	8.5407	1	2.3607	APAC_Consumer
10409	2014-12-30	Consumer	APAC	255.2850	2	-47.6550	APAC_Consumer
34455	2014-12-30	Consumer	APAC	10.8540	3	-6.6960	APAC_Consumer
13007	2014-12-31	Consumer	APAC	300.2400	3	84.0600	APAC_Consumer
30632	2014-12-31	Consumer	APAC	39.6000	3	6.6600	APAC_Consumer

5699 rows × 7 columns

APAC Consumer is the most profitable market segment as its total profit is maximum from all the other market segments and its Cov value is least.

```
# finding the total sum of 'Profit' for train data
sum_1=train.sum(axis=0, skipna=True)
sum_1.sort_values(ascending=True)
```

```
Market_Segment
Canada_Home Office       2764.95000
Canada_Corporate         3090.57000
EMEA_Home Office         5176.46700
EMEA_Corporate           7235.52900
Canada_Consumer          8282.43000
Africa_Home Office      13986.09000
EMEA_Consumer           17444.90100
Africa_Corporate        17893.14600
LATAM_Home Office       33118.33784
Africa_Consumer         33553.75500
US_Home Office          44620.69180
LATAM_Corporate         45191.75736
EU_Home Office          46092.54300
APAC_Home Office        57923.07120
US_Corporate            77849.88150
EU_Corporate            94583.70750
LATAM_Consumer          94612.45620
APAC_Corporate         107393.95620
US_Consumer            109356.92260
EU_Consumer            152355.71550
APAC_Consumer          177389.25060
```

Market_Segment	
APAC_Consumer	0.603633
EU_Consumer	0.655334
LATAM_Consumer	0.688935
EU_Corporate	0.697702
APAC_Corporate	0.740799
LATAM_Corporate	0.890930
US_Corporate	1.039660
APAC_Home Office	1.061530
US_Consumer	1.108571
EU_Home Office	1.128192
Canada_Corporate	1.219189
US_Home Office	1.231887
LATAM_Home Office	1.359984
Africa_Consumer	1.446661
Canada_Consumer	1.497032
Africa_Corporate	1.685008
Africa_Home Office	2.013987
Canada_Home Office	2.245148
EMEA_Consumer	2.749927
EMEA_Home Office	6.140222
EMEA_Corporate	6.861820

Data Preparation

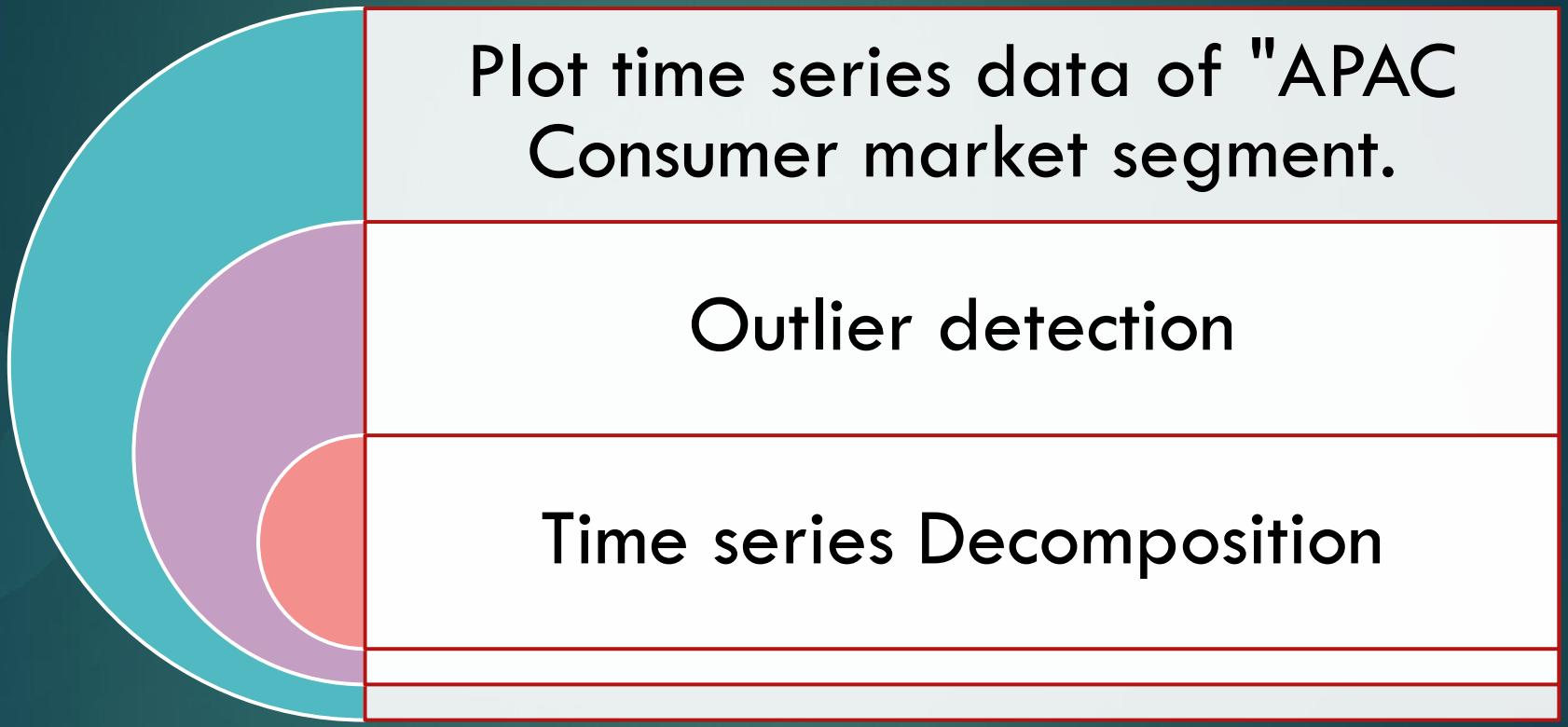
- Creating the monthly aggregated data for 'Order Date' and ‘Sales' of APAC_Consumer.

Order Date	Sales
2011-01-31	15711.7125
2011-02-28	12910.8588
2011-03-31	19472.5632
2011-04-30	15440.3046
2011-05-31	24348.9723
2011-06-30	27260.0196
2011-07-31	15842.8317
2011-08-31	22012.2366
2011-09-30	34613.1849
2011-10-31	36472.0470
2011-11-30	37722.6039
2011-12-31	37846.9911
2012-01-31	31280.8635
2012-02-29	24985.6881
2012-03-31	14241.1761
2012-04-30	20926.4721
2012-05-31	32608.6212
2012-06-30	39710.0352
2012-07-31	8389.7316
2012-08-31	48444.7977
2012-09-30	28193.2236
2012-10-31	56743.0833
2012-11-30	51967.0140

- Creating the monthly aggregated data for 'Order Date' and 'Quantity' of APAC_Consumer.

Order Date	Quantity
2011-01-31	214
2011-02-28	151
2011-03-31	283
2011-04-30	148
2011-05-31	244
2011-06-30	322
2011-07-31	212
2011-08-31	325
2011-09-30	391
2011-10-31	401
2011-11-30	504
2011-12-31	511
2012-01-31	279
2012-02-29	277
2012-03-31	132
2012-04-30	218
2012-05-31	434
2012-06-30	507
2012-07-31	156

Steps to follow for Analysing the Data



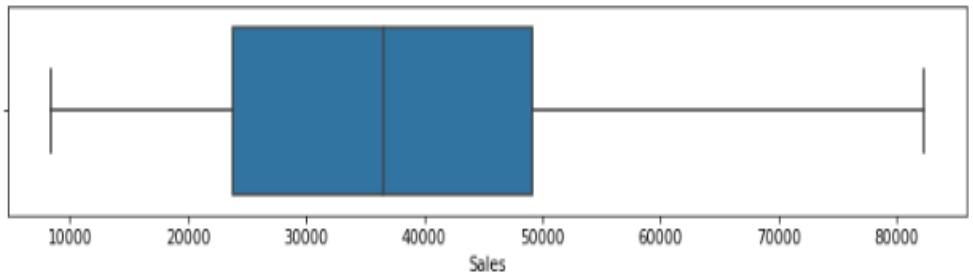
Plotting time series data of "APAC Consumer market segment for Sales and Quantity.

Inference:-The number of sales keeps increasing and thus, the data shows a upward and increasing trend. The data also shows a seasonal pattern which we can see from the above graph that after the month of june there is a sudden decrease in sales in month of july this clearly implies that the dataset has a seasonality component.

Inference:- The number of Quantity keeps increasing and thus, the data shows a upward and increasing trend. The data also shows a seasonal pattern which we can see from the above graph that after the month of june there is a sudden decrease in demand of Quantity in month of july this clearly implies that the dataset has a seasonality component.

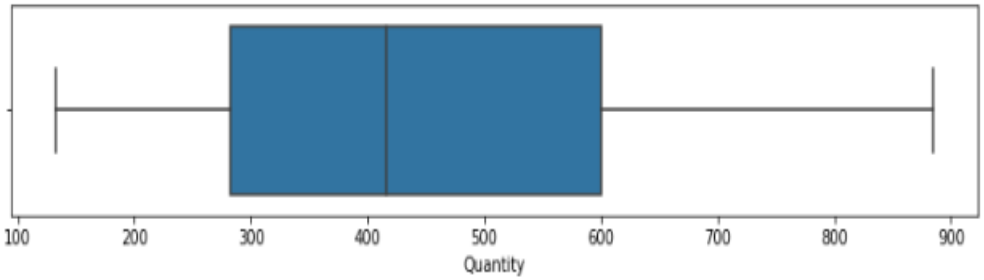
Outlier detection for Sales and Quantity of APAC Consumer market segment.

```
import seaborn as sns
fig= plt.subplots(figsize=(12,2))
ax= sns.boxplot(x=Sales_forecast['Sales'])
```



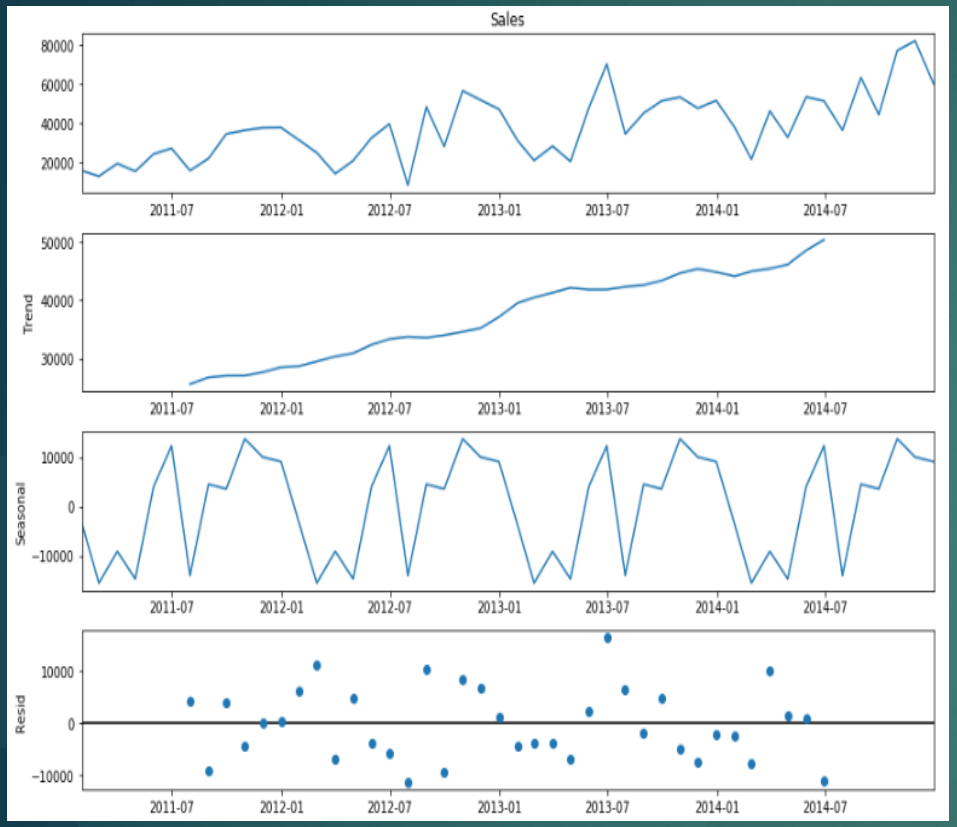
Inference:- No outliers are present.

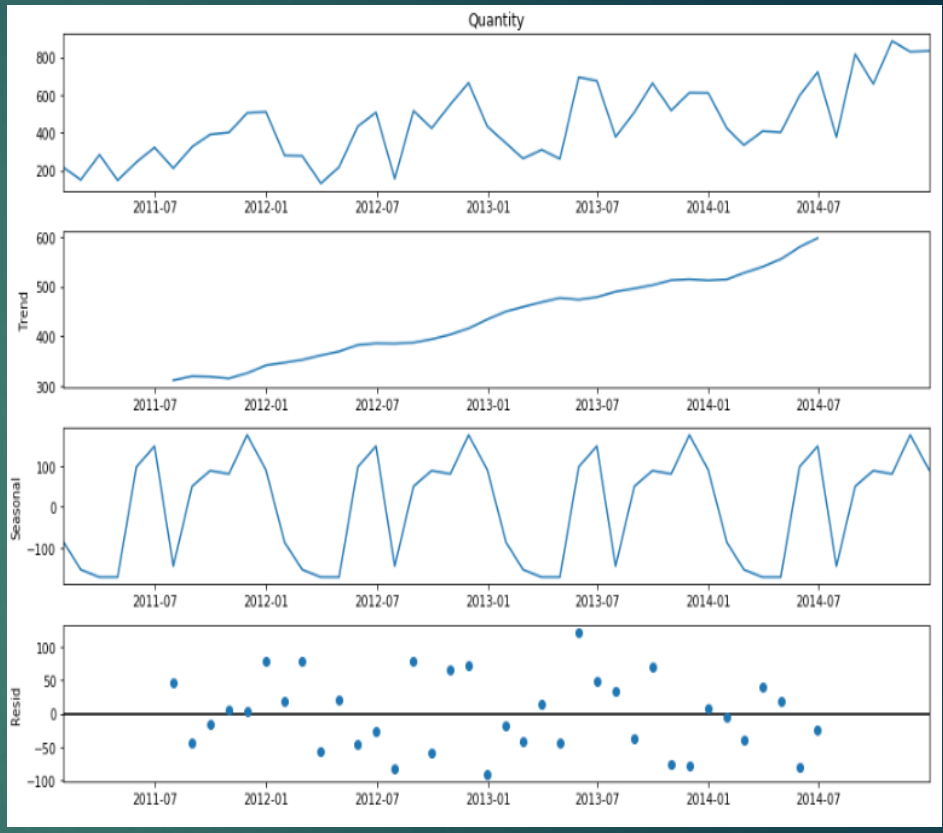
```
import seaborn as sns
fig= plt.subplots(figsize=(12,2))
ax= sns.boxplot(x=Quantity_forecast['Quantity'])
```



Inference:- No outliers are present.

Time series Decomposition Additive Seasonal Decomposition

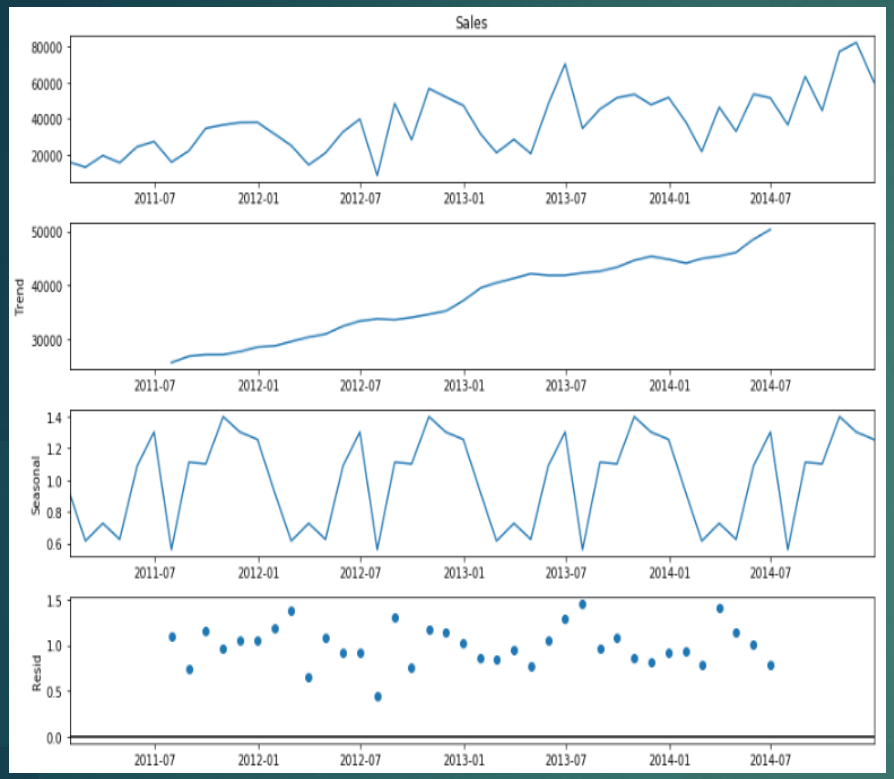


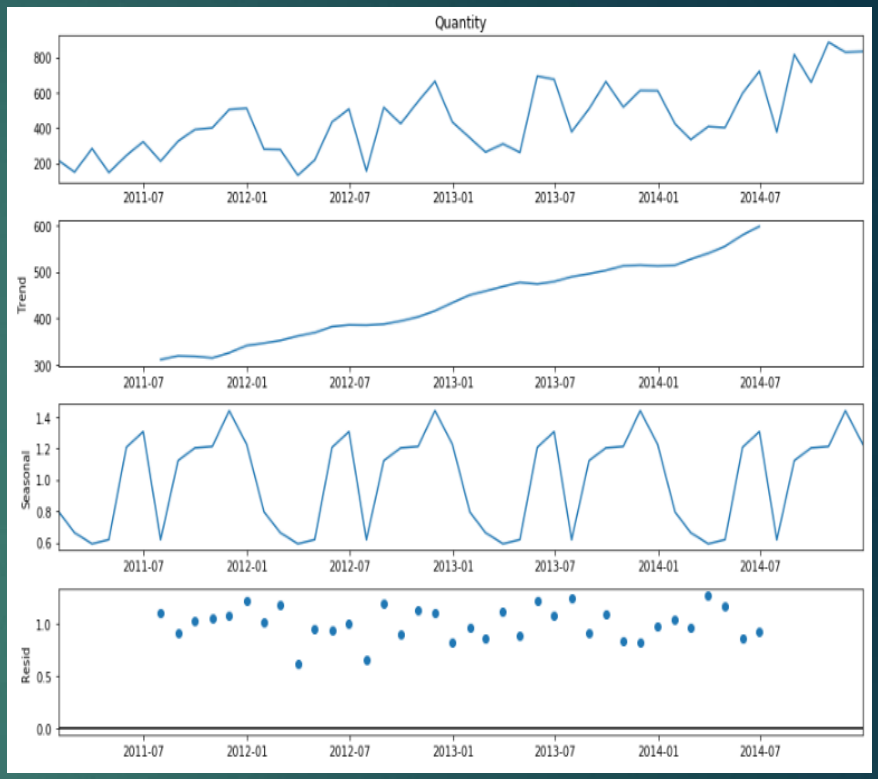


Inference:- Additive Seasonal Decomposition clearly shows the series has a clear upward trend and seasonal component.

Inference:- Additive Seasonal Decomposition clearly shows the series has a clear upward trend and seasonal component.

Time series Decomposition Multiplicative Seasonal Decomposition





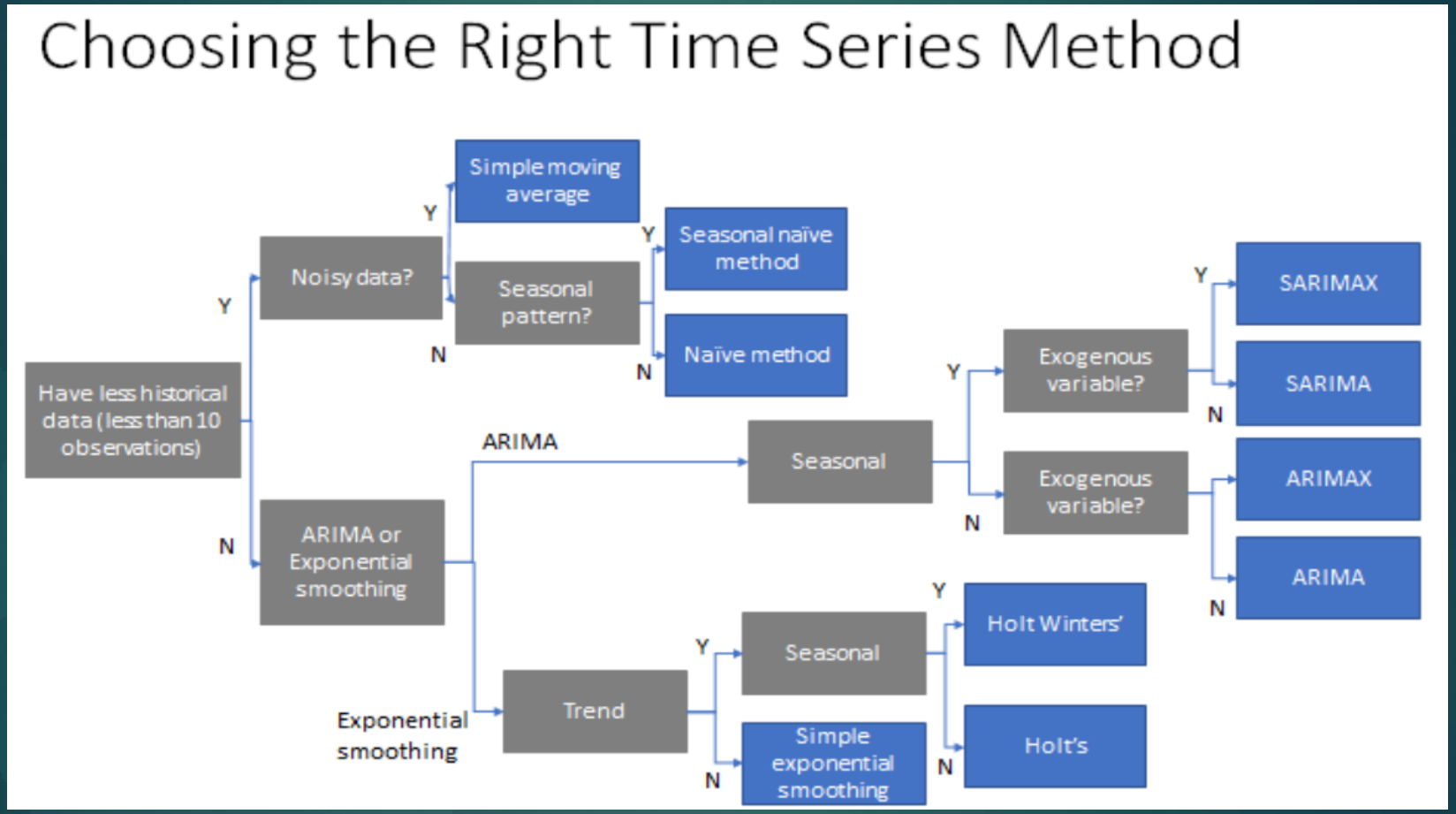
Inference:- Multiplicative Seasonal Decomposition clearly shows the series has a clear upward trend and seasonal component.

Inference:- Multiplicative Seasonal Decomposition clearly shows the series has a clear upward trend and seasonal component.

Building of Models

- Time series data was divided into train(1-42 month) & test sets(43-48 month).
- Models are prepared by applying the smoothing techniques and the ARIMA set of techniques.
- For ARIMA set of techniques, the data is converted into stationary time series by using box-cox transformation and differencing method.
- All models were evaluated using Mean Absolute Percentage Error(MAPE).

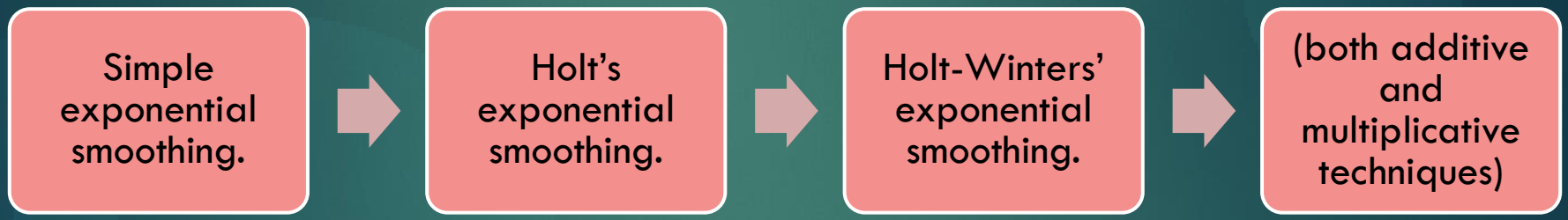
The optimum technique from the flow chart that might work best for both the sales and the quantity forecasts



Inference:- To capture the level, the trend and seasonality, the Holt-Winters' exponential smoothing technique / SARIMA works best.

Building all the models

First, starting by applying the smoothing techniques:-

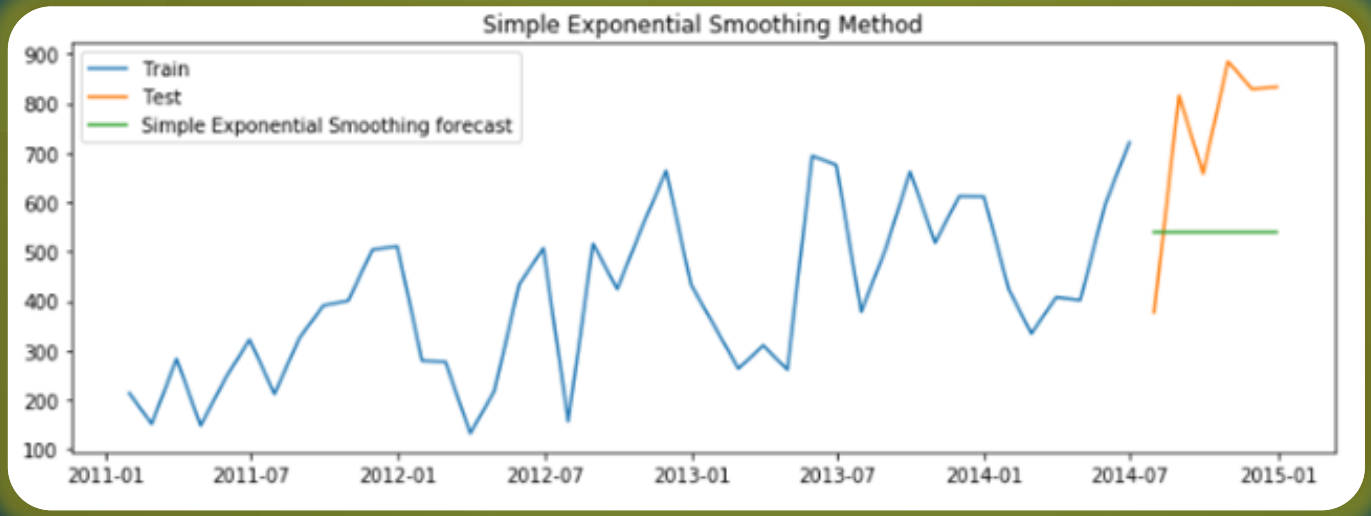


Inference:- On plotting the Sales and Quantity data set, I observed that Sales and Quantity has a upward, increasing trend and seasonal component, and thus **Holt-Winters' and SARIMA method** will work best for this data.

Simple Exponential Smoothing Method

Sales It helps us in capturing the level of time series data

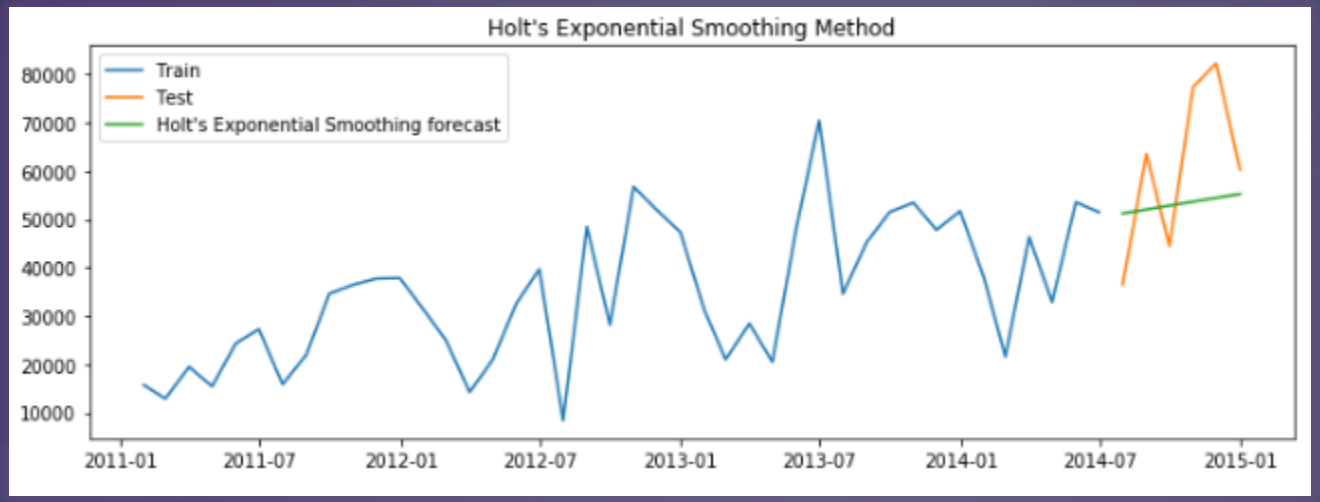
Quantity



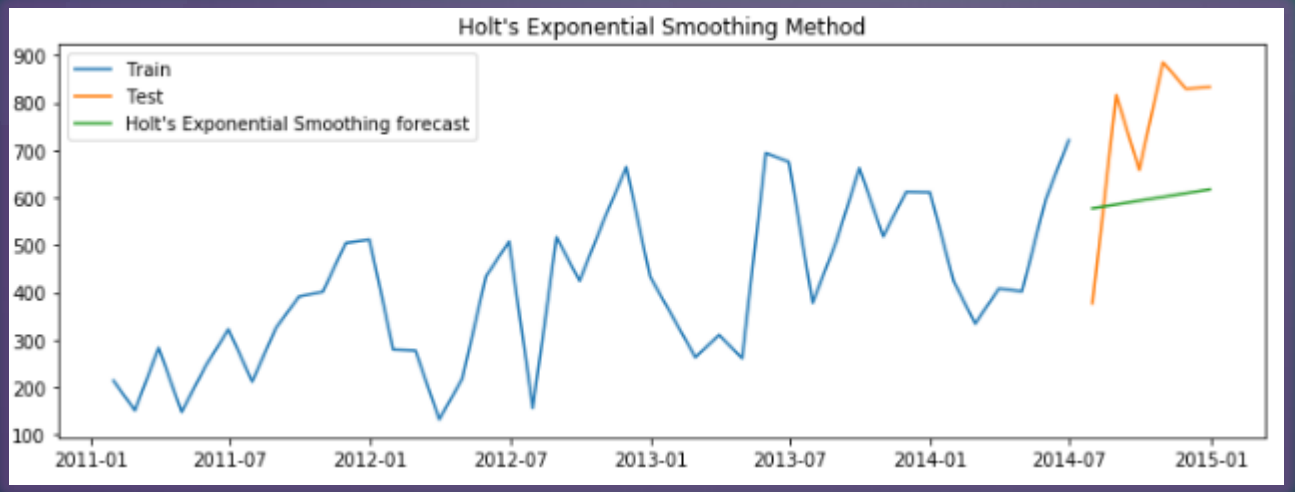
Holt's exponential smoothing Method

Technique which forecasts based on level, trend of a time series

Sales



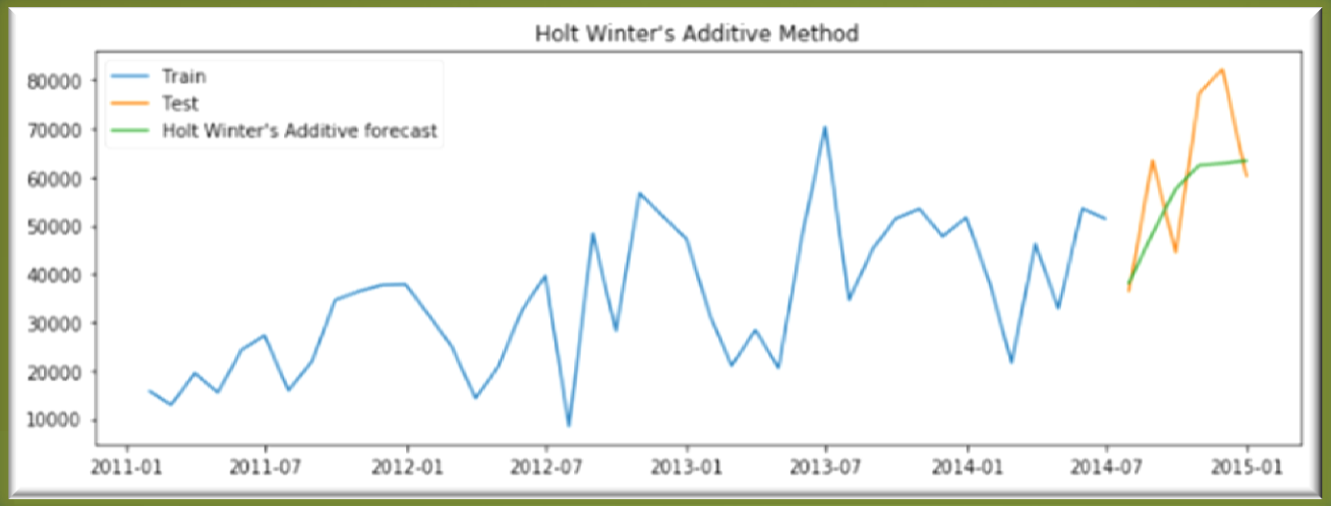
Quantity



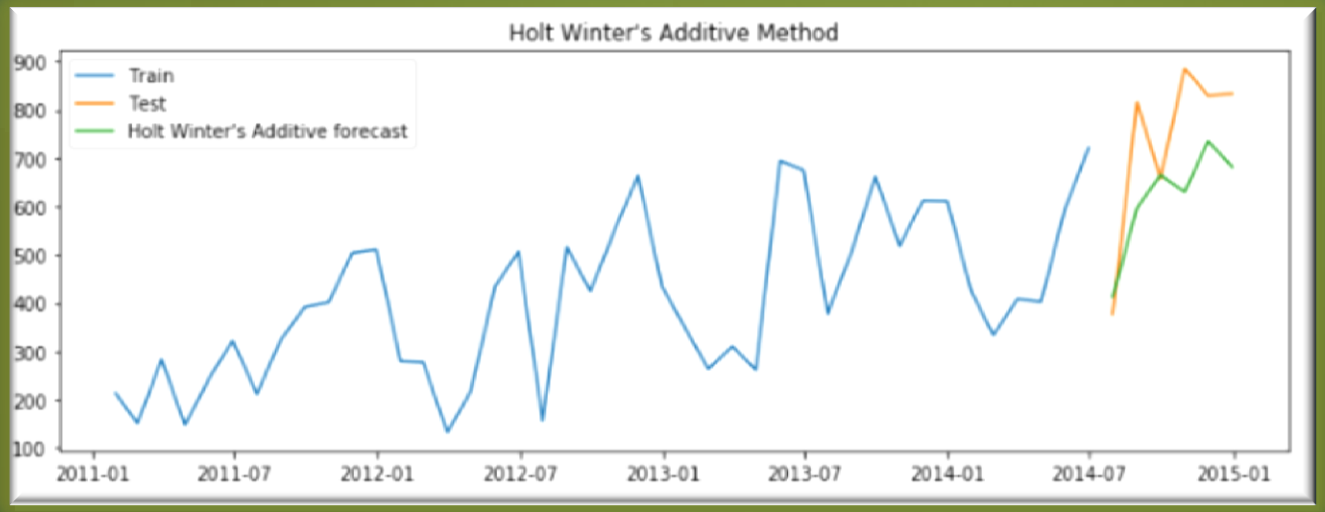
Holt Winters's Additive Exponential Smoothing Method

Technique which forecasts based on level, trend and seasonality of a time series

Sales

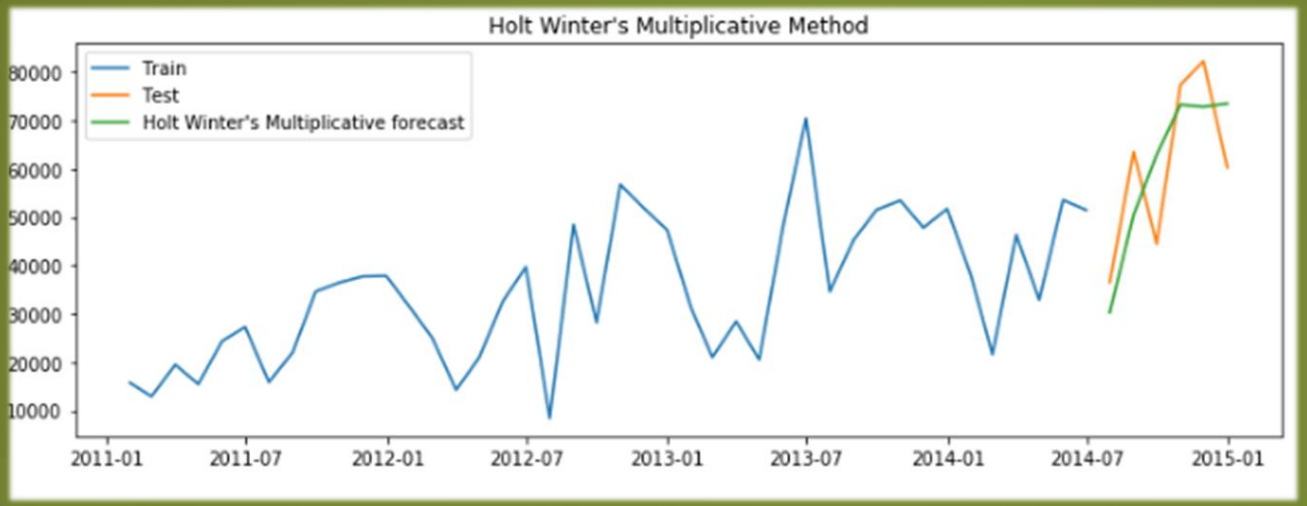


Quantity

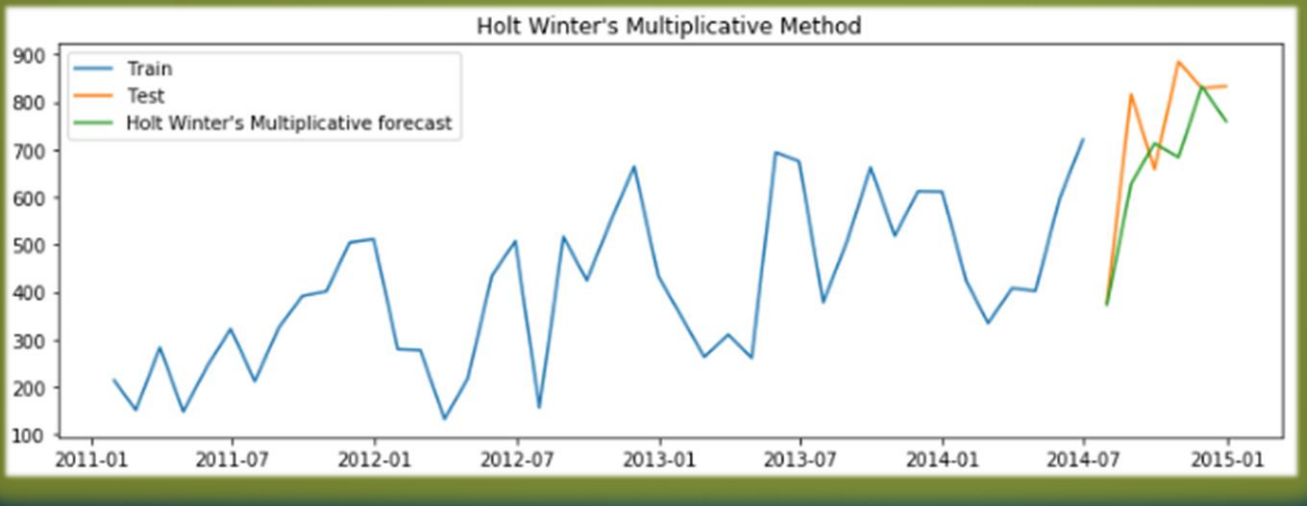


Holt Winters's Multiplicative Exponential Smoothing Method

Sales



Quantity



Applying the ARIMA set of techniques (Auto Regressive Models):-

1) AR model

2) MA model

3) ARMA model

4) ARIMA model

5) SARIMA model

ARIMAX and SARIMAX model (can not be applied, as there is no exogenous variable in the data).

Inference:- On plotting the Sales and Quantity data set, I observed that Sales and Quantity has a upward, increasing trend and seasonal component, and thus **Holt-Winters' and SARIMA method** will work best for this data.

Performed Augmented Dickey Fuller (ADF) test for Sales and Quantity datas.

Augmented Dickey Fuller (ADF) test

- Null Hypothesis (H0): The series is not stationary
- p-value > 0.05
- Alternate Hypothesis (H1): The series is stationary
- p-value ≤ 0.05

- p-value > 0.05: Fail to reject the null hypothesis (H0).
- p-value ≤ 0.05: Reject the null hypothesis (H0).

```
from statsmodels.tsa.stattools import adfuller
adf_test= adfuller(Sales_forecast['Sales'])

print('ADF statistic: %f' % adf_test[0])
print('Critical value @ 0.05: % .2f' % adf_test[4]['5%'])
print('p-value: %f' % adf_test[1])
# p-value is more than 0.05. We fail to reject the null hypothesis. So, it is not a stationary time series.
```

```
ADF statistic: -2.220857
Critical value @ 0.05: -2.93
p-value: 0.198763
```

Inference:- Sales data is not a stationary time series.

Augmented Dickey Fuller (ADF) test

- Null Hypothesis (H0): The series is not stationary
- p-value > 0.05
- Alternate Hypothesis (H1): The series is stationary
- p-value ≤ 0.05

- p-value > 0.05: Fail to reject the null hypothesis (H0).
- p-value ≤ 0.05: Reject the null hypothesis (H0).

```
from statsmodels.tsa.stattools import adfuller
adf_test= adfuller(Quantity_forecast['Quantity'])

print('ADF statistic: %f' % adf_test[0])
print('Critical value @ 0.05: % .2f' % adf_test[4]['5%'])
print('p-value: %f' % adf_test[1])
# p-value is more than 0.05. We fail to reject the null hypothesis. So, it is not a stationary time series.
```

```
ADF statistic: 0.293145
Critical value @ 0.05: -2.94
p-value: 0.977028
```

Inference:- Quantity data is not a stationary time series.

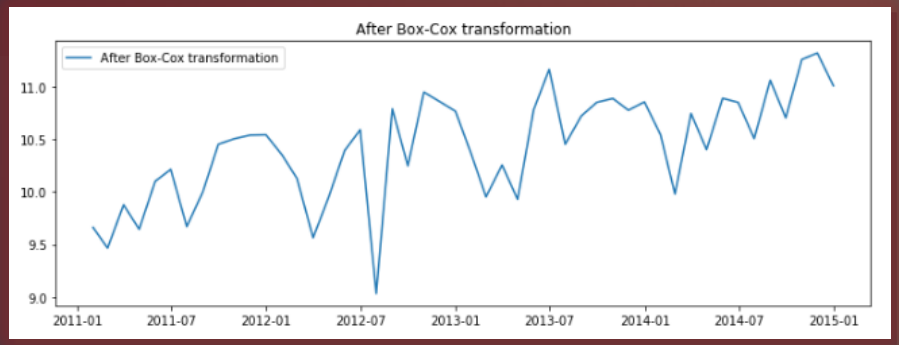
Auto Regressive Models

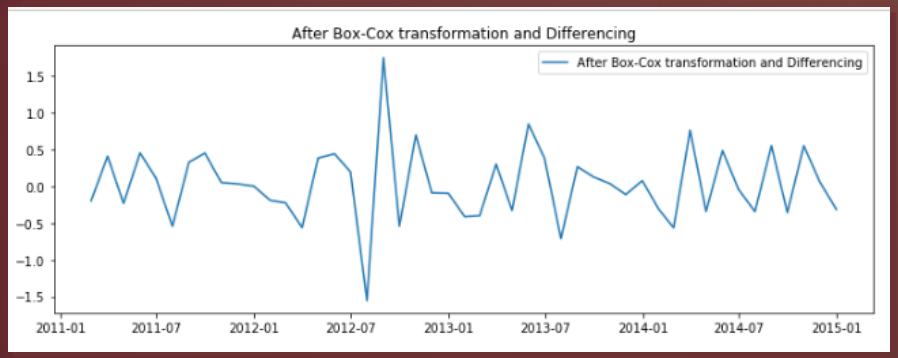
There are two fundamental assumptions to build an autoregressive model. They are –

1) Stationarity :- Stationary processes are easier to analyze and model because their statistical properties remain constant over time. There will be no trend, seasonality and cyclicity in the series. In other words, if the past observations and future observations follow the same statistical properties i.e. there are no change in mean, variance and covariance then the future observation can be easily predicted.

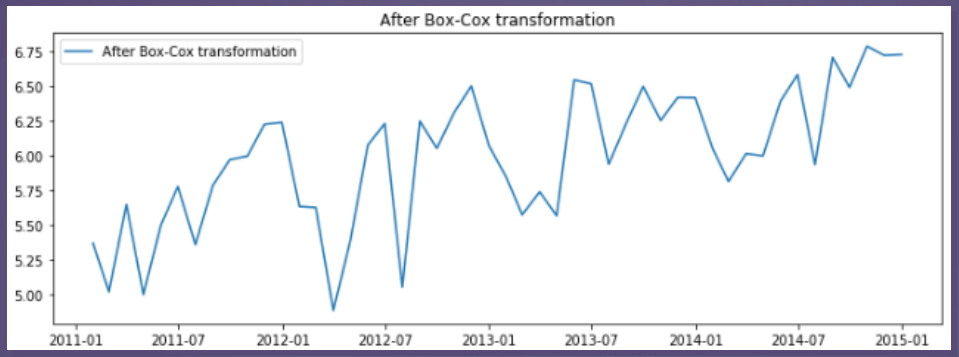
2) Autocorrelation:- Autocorrelation is capturing the relationship between observations yt at time t and yt−k at time k time period before t. In simpler words, autocorrelation helps us to know how a variable is influenced by its own lagged values.

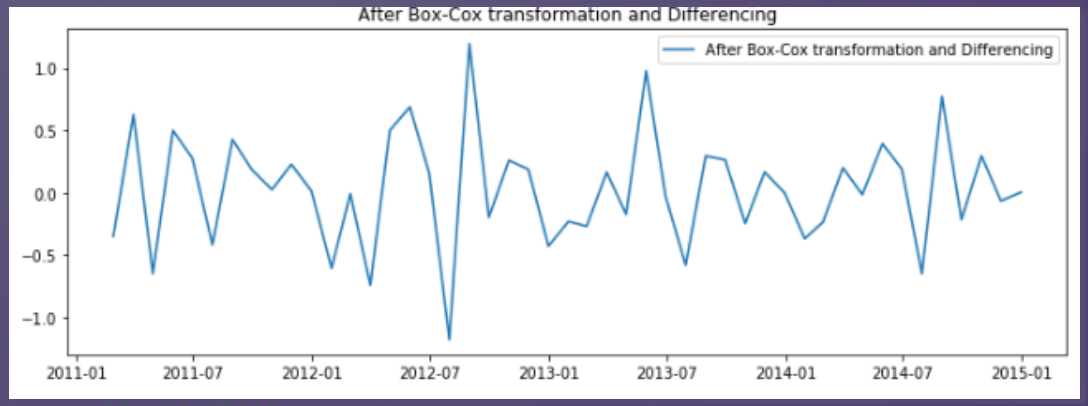
Converting Sales data into stationary time series by box-cox transformation and differencing method.





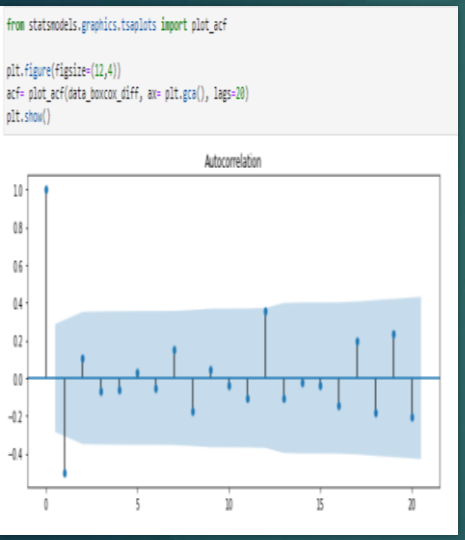
Converting Quantity data into stationary time series by box-cox transformation and differencing method.



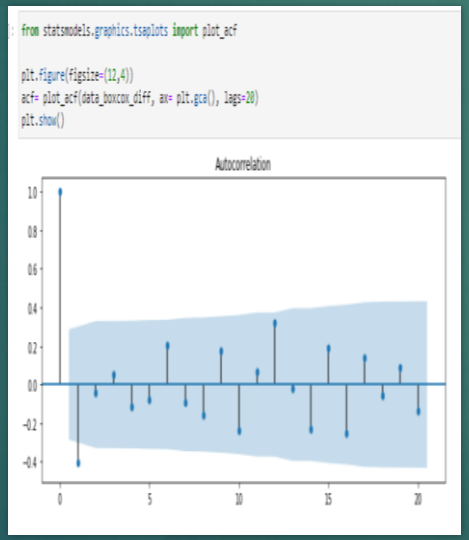


Autocorrelation :The autocorrelation function tells about the correlation between an observation with its lagged values.

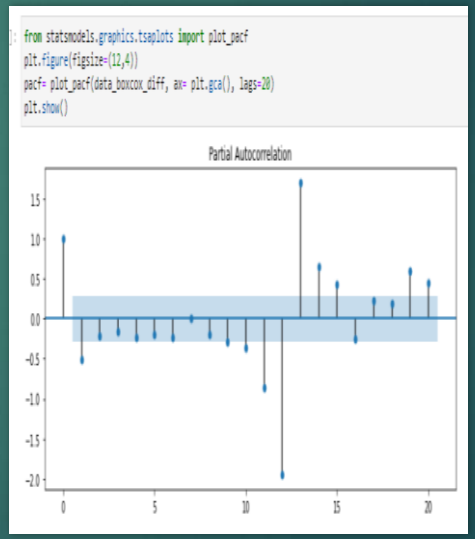
ACF of Sales



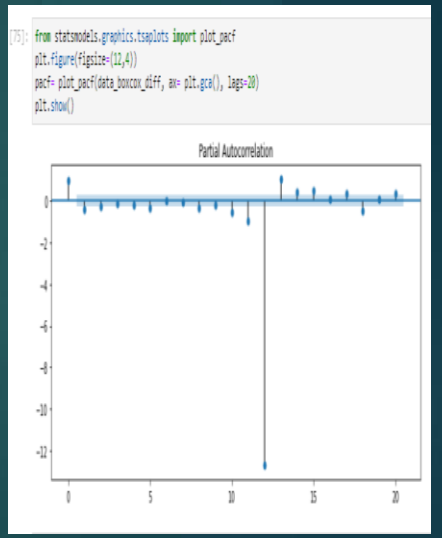
ACF of Quantity



PACF for sales

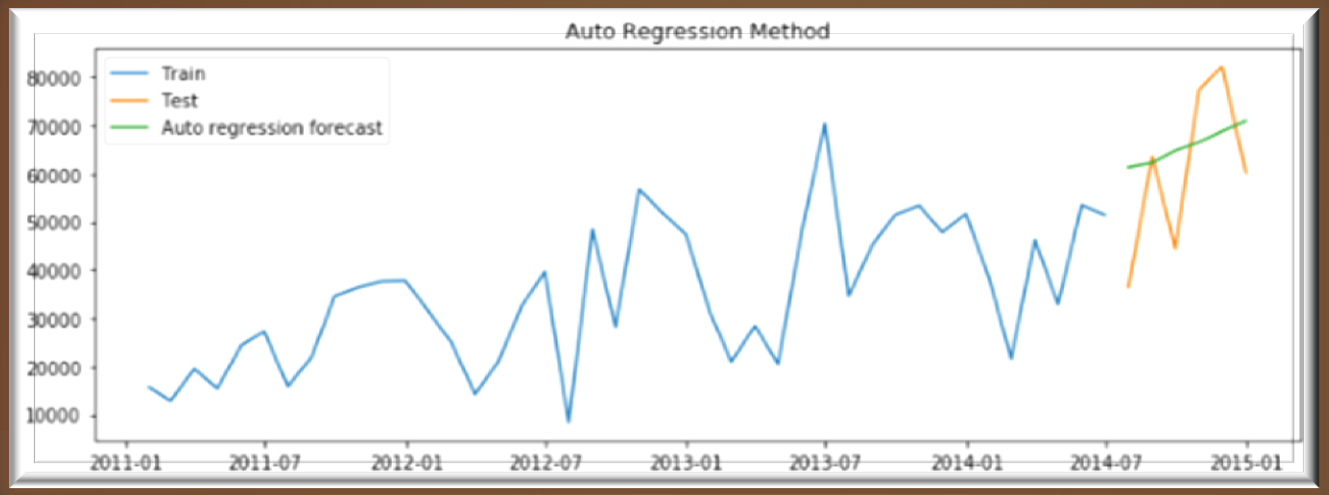


PACF of Quality

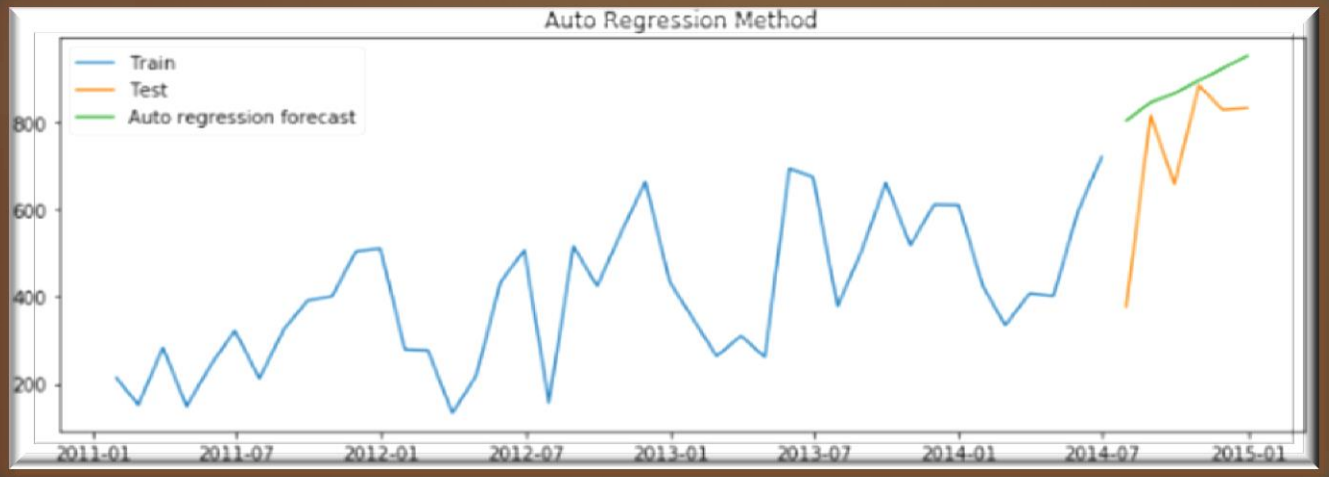


Auto Regression Method (AR) order=(1,0,0)

Sales

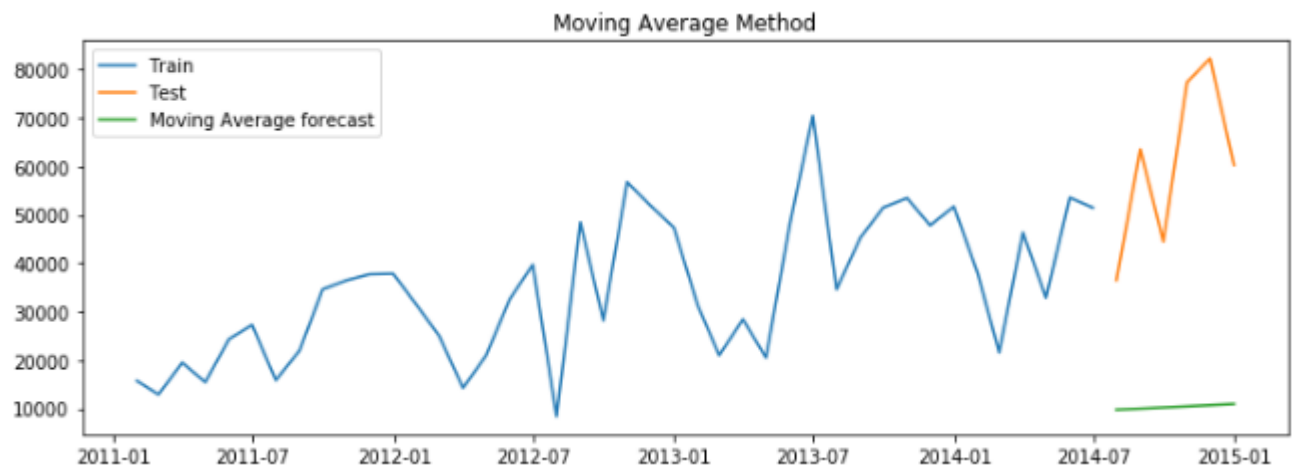


Quantity

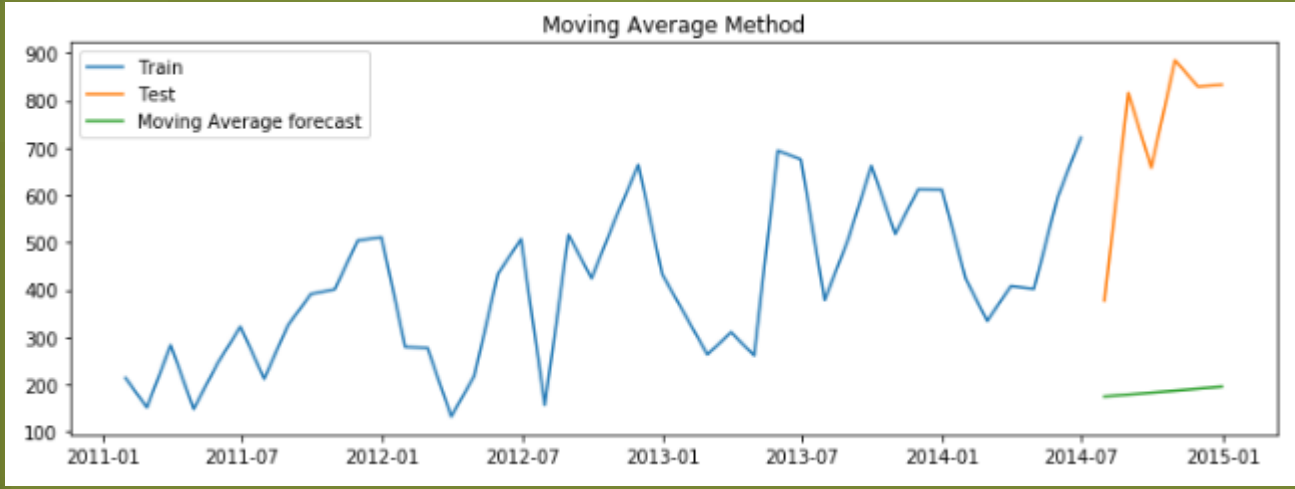


Moving Average Method (MA) order=(0,0,1)

Sales

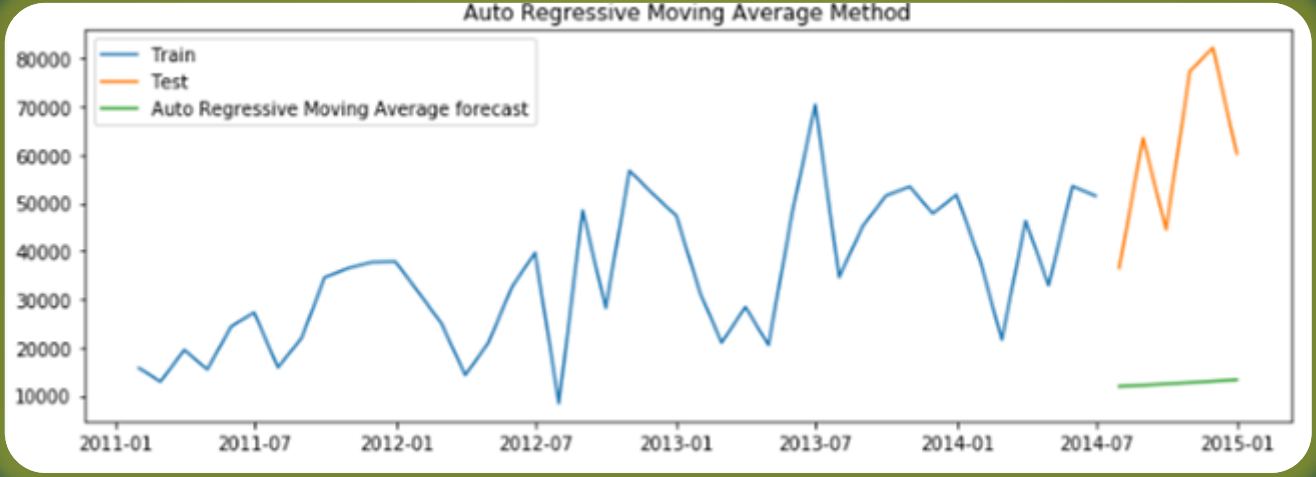


Quantity

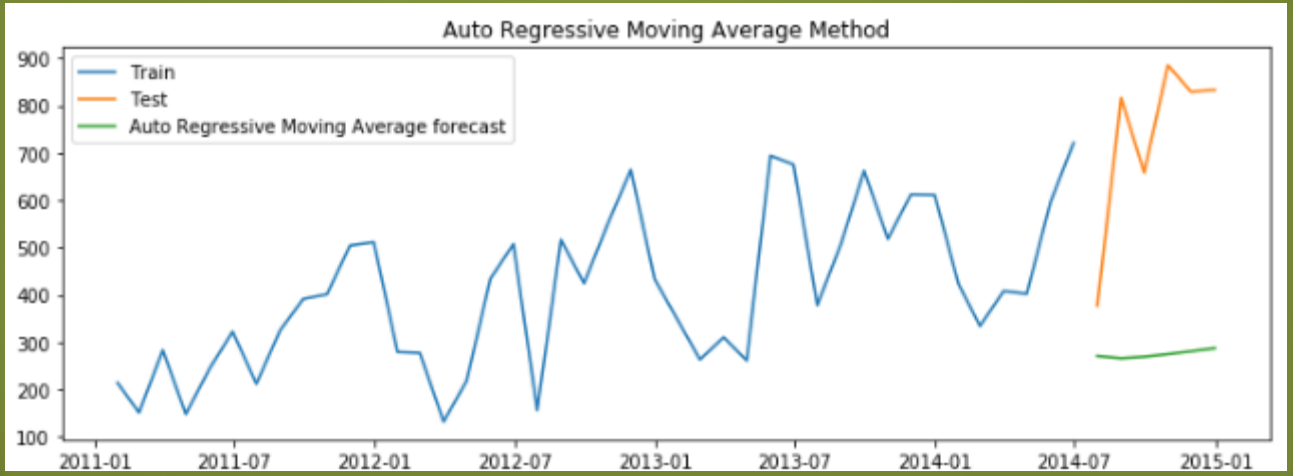


Auto Regressive Moving Average Method (ARMA) order=(1,0,1)

Sales

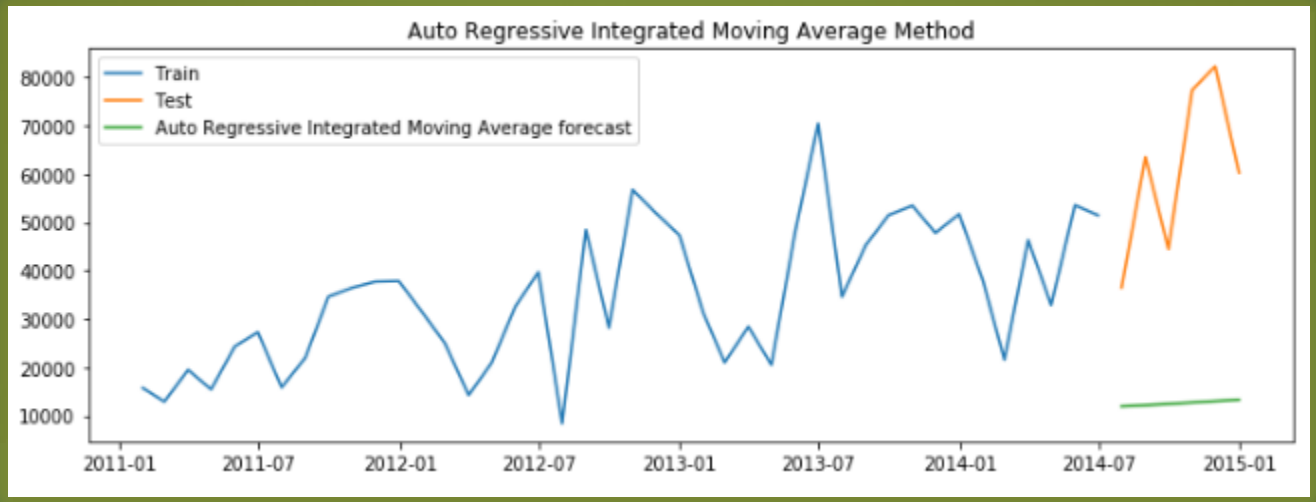


Quantity

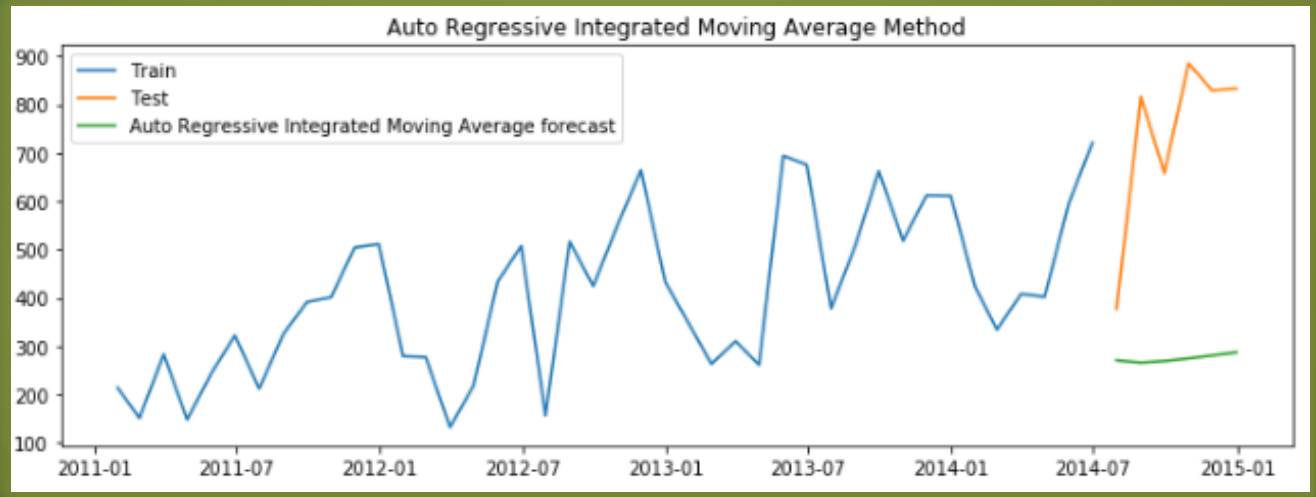


Auto Regressive Integrated Moving Average Method (ARIMA) order=(1,1,1)

Sales

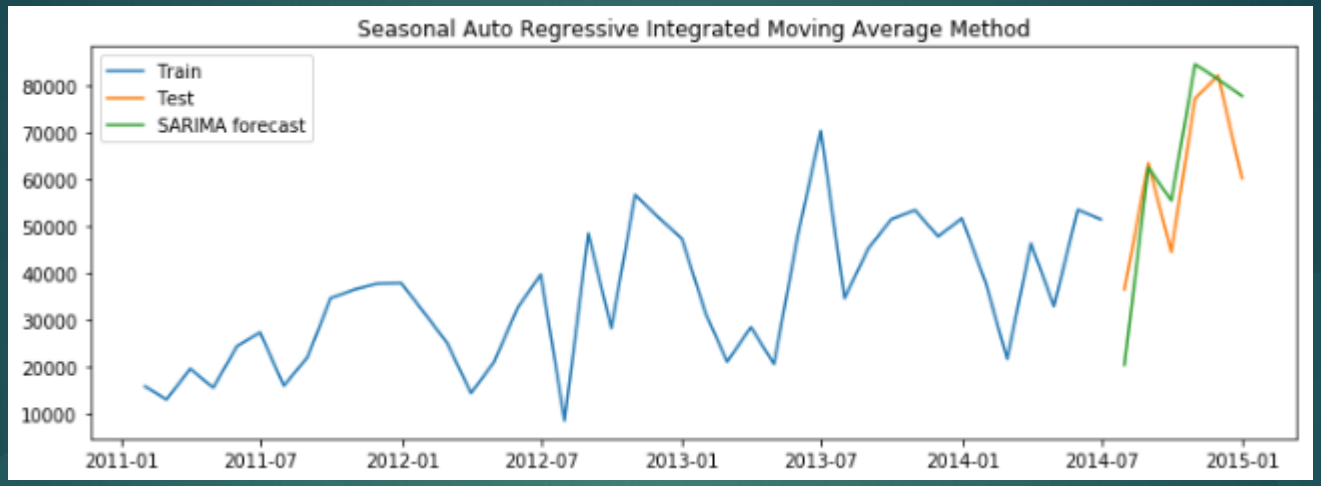


Quantity

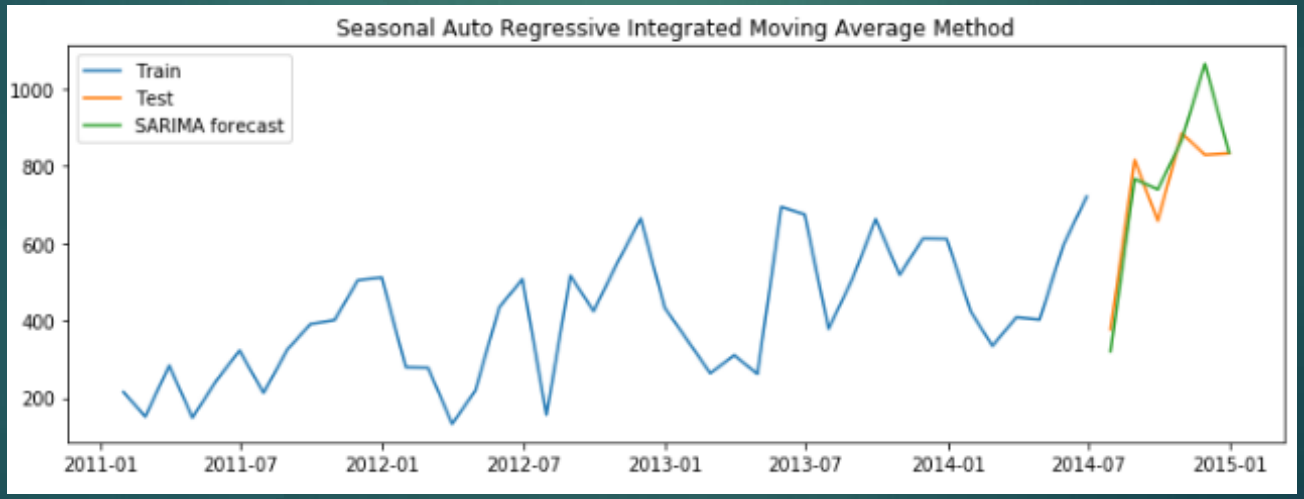


Seasonal Auto Regressive Integrated Moving Average Method (SARIMA) order=(1,1,1) seasonal_order=(1,1,1,12)

Sales



Quantity



Mean Absolute Percentage Error (MAPE)

Sales

Method	MAPE	RMSE
Naive Method	26.86	18774.05
Simple Average Method	38.18	30846.00
Simple Moving Average Method	28.15	23383.65
Simple Exponential Smoothing Method	27.73	22992.97
Holt's Exponential Smoothing Method	25.00	17228.59
Holt Winter's Additive Method	17.61	12971.01
Holt Winter's Multiplicative Method	19.62	11753.42
Auto Regressive(AR) Method	27.27	15505.02
Moving Average (MA) Method	81.64	52903.35
Auto Regressive Moving Average (ARMA) Method	77.66	50758.27
Auto Regressive Integrated Moving Average (ARI...	77.66	50758.27
SARIMA Method	18.38	11179.94

Quantity

Method	MAPE	RMSE
Naive Method	26.24	174.37
Simple Average Method	42.16	371.15
Simple Moving Average Method	35.55	279.36
Simple Exponential Smoothing Method	34.15	261.71
Holt's Exponential Smoothing Method	29.29	213.30
Holt Winter's Additive Method	15.98	156.24
Holt Winter's Multiplicative Method	10.73	118.69
Auto Regressive(AR) Method	29.30	204.28
Moving Average (MA) Method	72.76	573.80
Auto Regressive Moving Average (ARMA) Method	59.24	489.40
Auto Regressive Integrated Moving Average (ARI...	59.24	489.40
SARIMA Method	10.69	106.60

Conclusion:- The **Holt Winter's (additive and multiplicative)** and **SARIMA** method have least MAPE values among all the other methods done above.
This technique forecasts based on level, trend and seasonality of a time series.

Forecast of test set (6 months)

Sales

Order Date	Sales	sarima	hwa_forecast	hwm_forecast
2014-07-31	36524.3028	20334.821581	38120.571471	30278.415968
2014-08-31	63521.7729	62601.380828	48451.426795	50494.663578
2014-09-30	44477.2662	55512.001802	57615.965549	62901.684796
2014-10-31	77379.8286	84721.838906	62518.929420	73280.128482
2014-11-30	82286.3583	81488.291897	62885.180755	72836.867938
2014-12-31	60292.1310	77918.594155	63445.198577	73505.060185

Quantity

Order Date	Quantity	sarima	hwa_forecast	hwm_forecast
2014-07-31	377	320.219248	412.603446	373.227992
2014-08-31	816	766.390125	596.452647	627.627281
2014-09-30	658	739.071363	665.429487	712.543317
2014-10-31	885	867.578890	630.085274	683.503273
2014-11-30	829	1064.836590	734.348021	832.421480
2014-12-31	833	835.110388	681.357068	759.278085

Conclusion:- The method whose forecast is able to predict the sales and Quantity closer to the actual values are **Holt Winter's (additive and multiplicative)** and **SARIMA** method.
These both method have least MAPE values among all the other methods done above.
This technique forecasts based on level, trend and seasonality of a time series.

inference

- Consumer Segment of APAC is the most consistent in terms of profit
- Sales and Quantity in both the markets seems to show seasonal pattern.
- Slightly increasing trend is present in both sales and quantity
- We created total 12 forecasting models for the most profitable market_segment (APAC consumer) out of which Holt Winter's (additive and multiplicative) and SARIMA method is the one whose forecast is able to predict the sales and Quantity closer to the actual values and there MAPE values is the least among all the other methods.
- Both above mentioned techniques forecasts based on level, trend and seasonality of a time series.
- The best technique in smoothing techniques is Holt Winter's (additive and multiplicative)
- The best technique in ARIMA set of techniques is Seasonal Auto Regressive Integrated Moving Average (SARIMA)